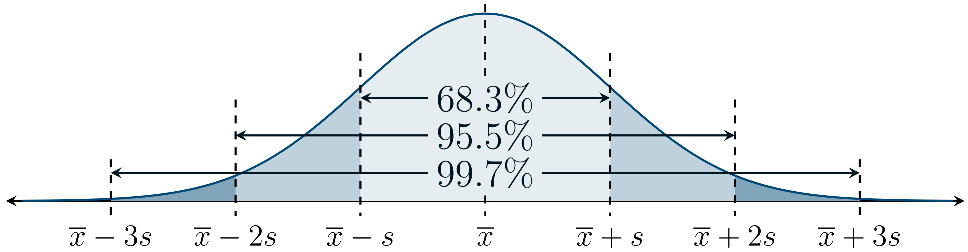
3.2: What's Unusual? The Empirical Rule and z-Scores

Definition.
The **Empirical Rule** is a guideline for how the standard deviation measures variability. If the distribution is unimodal and symmetric, then

- Approximately 68% of the observations will be within one standard deviation of the mean.
- Approximately 95% of the observations will be within two standard deviations of the mean.
- Nearly all of the observations will be within three standard deviations of the mean.



Example. Recall the data set of gas prices from before.

\$2.19	\$2.19	\$2.39	\$2.19
\$2.24	\$2.39	\$2.27	\$2.29
\$2.17	\$2.29	\$2.30	\$2.29

Recall that the mean gas price was \$2.2666... with a standard deviation of \$0.074. If this is representative of a larger data set, then...

- approximately 68% of the prices would fall between \$2.19 and \$2.34,
- approximately 95% of the prices would fall between \$2.12 and \$2.41, and
- nearly all of the prices would fall between \$2.04 and \$2.49.

Example. The mean daily high temperature in San Francisco is $65°F$ with a standard deviation of $8°F$.

- Find the temperature ranges for 68%, 95%, and 99.7% of the data.

$$68\% :$$
$$(\bar{x} - 1s, \bar{x} + 1s) = (65 - 1 \cdot 8, 65 + 1 \cdot 8) = (57, 73)$$

$$95\% :$$
$$(\bar{x} - 2s, \bar{x} + 2s) = (65 - 2 \cdot 8, 65 + 2 \cdot 8) = (49, 81)$$

$$99.7\% :$$
$$(\bar{x} - 3s, \bar{x} + 3s) = (65 - 3 \cdot 8, 65 + 3 \cdot 8) = (41, 89)$$

- By the Empirical rule, observations 2 or more standard deviations away from the mean are considered unusual. Is it unusual to have a day when the maximum temperature is colder than $49°F$ in San Francisco?

Since 49 is 2 standard deviations away from the mean, temperatures lower than this are considered unusual

Example. Suppose that after computing the mean $\bar{x}$ and standard deviation s, we conclude from the empirical rule that approximately 68% of our data lies between 6.5 and 14.78.

- Find the mean $\bar{x}$ and standard deviation s

$$\bar{x} = \frac{6.50 + 14.78}{2} = 10.64$$

$$s = 14.78 - 10.64 = 4.14$$

- Use the empirical rule to find the bounds that contain approximately 95% of the data.

$$\bar{x} - 2 \cdot s = 10.64 - 2 \cdot 4.14 = 2.36$$
$$\bar{x} + 2 \cdot s = 10.64 + 2 \cdot 4.14 = 18.92$$

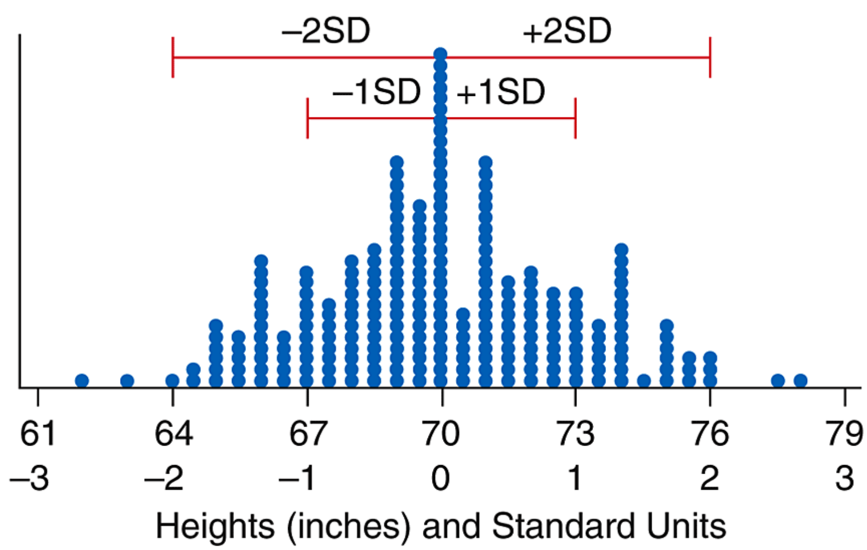
Definition.
A z**-score** measures how many standard deviations an observed data value, x, is from the mean $\overline{x}$:

$$z = \frac{x - \overline{x}}{s}$$

Example. The dotplot below shows the heights (in inches) of 247 men. The average height is 70 inches, and the standard deviation is 3 inches. How many men have z-scores...

- greater than 2? 2
- less than -2? 2
- What is the z-score of a man who is 75 inches tall?

$$z - \text{score} = \frac{75 - 70}{3} = 1.66....$$



Example. Maria scored 80 out of 100 on her first stats exam in a course and 85 out of 100 on her second stats exam. On the first exam, the mean was 70 and the standard deviation was 10. On the second exam, the mean was 80 and the standard deviation was 5.

On which exam did Maria perform better when compared to the whole class?

$$\begin{aligned} x_1 &= 80 \\ \bar{x}_1 &= 70 \\ s_1 &= 10 \end{aligned} \qquad\qquad \begin{aligned} x_2 &= 85 \\ \bar{x}_2 &= 80 \\ s_2 &= 5 \end{aligned}$$

$$\begin{aligned} \Rightarrow z_1 &= \tfrac{x-\bar{x}}{s} \\ &= \tfrac{80-70}{10} = 1 \end{aligned} \qquad\qquad \begin{aligned} \Rightarrow z_2 &= \tfrac{x-\bar{x}}{s} \\ &= \tfrac{85-80}{5} = 1 \end{aligned}$$

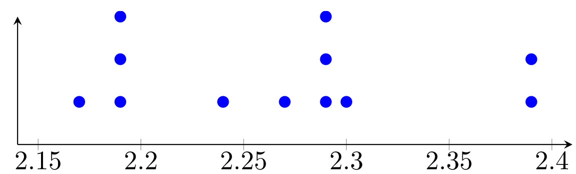
Maria's z-scores were both 1, indicating she was 1 standard deviation above the mean. In comparison to her class, she did equally well on both exams

3.3: Summaries for Skewed Distributions

> **Definition.**
> The **median** of a sample of data is the middle value when the data is sorted from smallest to largest. If the set contains an
>
> - odd number of observed values, the median is the middle observed value.
> - even number of observed values, the median is the average of the two middle observed values.
>
> The median is the preferred measure of center when the data is skewed since about 50% of the observations lie below and above the median.

Example. The prices of 1 gallon of regular gas at 12 service stations near the downtown area of Austin, TX, were as follows one winter day in 2018:

\$2.19	\$2.19	\$2.39	\$2.19
\$2.24	\$2.39	\$2.27	\$2.29
\$2.17	\$2.29	\$2.30	\$2.29



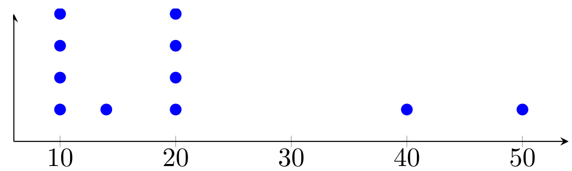
Find the median price for a gallon of gas and interpret the value.

2.39
2.39
2.30
2.29
2.29
2.29
2.27
2.24
2.19
2.19
2.19
2.17

$$\text{median} = \frac{2.27 + 2.29}{2} = 2.28$$

Example. Below are the percentages of fat for some brands of sliced turkey:

$14, 10, 20, 20, 40, 20, 10, 10, 20, 50, 10$



Find the median percentage of fat and interpret the value.

10, 10, 10, 10, 14, $\boxed{20,}$ 20, 20, 20, 40, 50

The typical percentage fat is 20%

Definition.

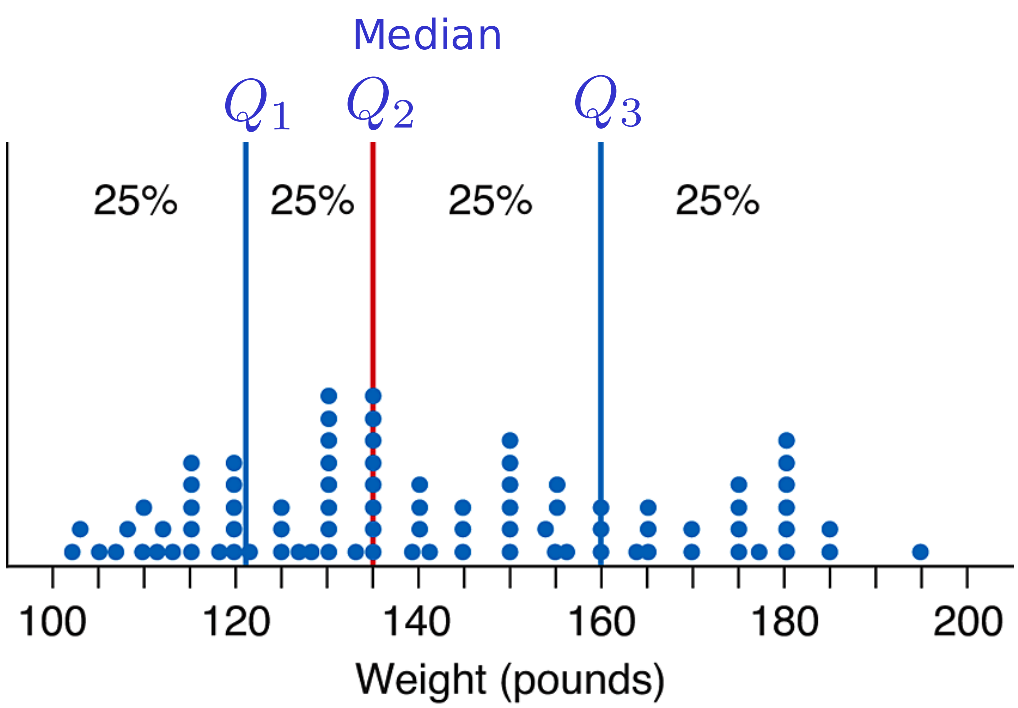
- The **range** is the difference between the maximum and minimum values:

$$\text{Range} = \text{maximum} - \text{minimum}$$

- The **quartiles** divide the data into quarters.
- The **interquartile range (IQR)** indicates approximately how much space the middle 50% of the data occupy.

Example. The dotplot below shows the distribution of weights for a class of introductory statistics students.

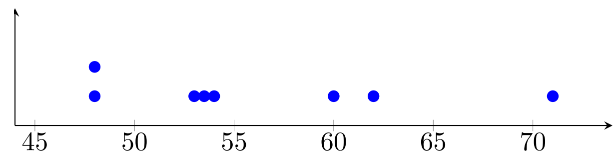
- Label each line
- Compute the IQR $IQR = Q_3 - Q_1 = 160 - 121 = 39$



Example (Computing quartiles by hand)**.** A group of eight children have the following heights (in inches):

Q_1 Q_2 Q_3

48.0, 48.0, 53.0, 53.5, 54.0, 60.0, 62.0, 71.0



Find the following:

- The median, which is also referred to as Q2

$$Q_1 = \frac{53.5 + 54}{2} = 53.75$$

- The first quartile (Q1), which is the median of the lower half of the sorted data

$$Q_1 = \frac{48 + 53}{2} = 50.5$$

- The third quartile (Q3), which is the median of the upper half of the sorted data

$$Q_3 = \frac{60 + 62}{2} = 61$$

- Compute the IQR

IQR=61-50.5=10.5